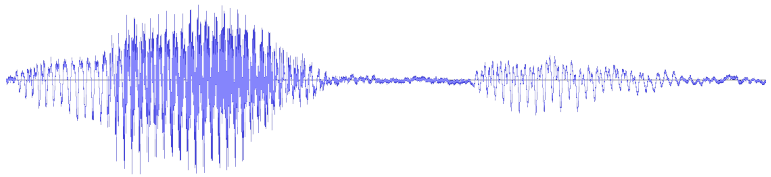
Multi-modal Speech Representation Learning

Carlos Schmidt

05.12.2023



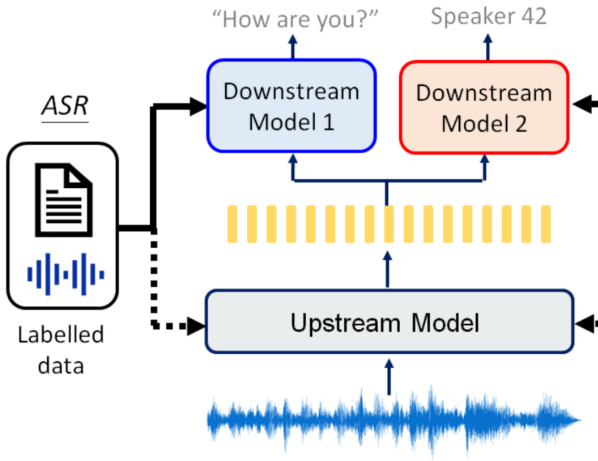
Lorem ipsum dolor sit amet, consectetur adipisici elit, ...

Intro

Learning paradigms
Single-mode approaches
Multi-modal approaches

Speech Representation Learning

- Building a single model for various downstream tasks
- Related to Textual Word Embeddings
- Training speech representation models
  - Train upstream model
  - Finetune to downstream task



Qualities of Speech Representations [

- Disentangled speaker identity, style, emotion, …
- Noise invariance
- Hierarchy w.r.t. low level vs. high level features
  - e.g., speaker identification vs. translation task

History of Speech Representation Learning

- Earliest approaches [1]:
  - Clustering approaches
    - k-means, GMMs
  - Hidden Markov Models: Allow processing of continuous speech
- Currently: Pretext task optimization (pre-training)
  - Learn Representation by solving “pretext” task
    - e.g., predicting masked tokens in sequence

Intro

Learning paradigms

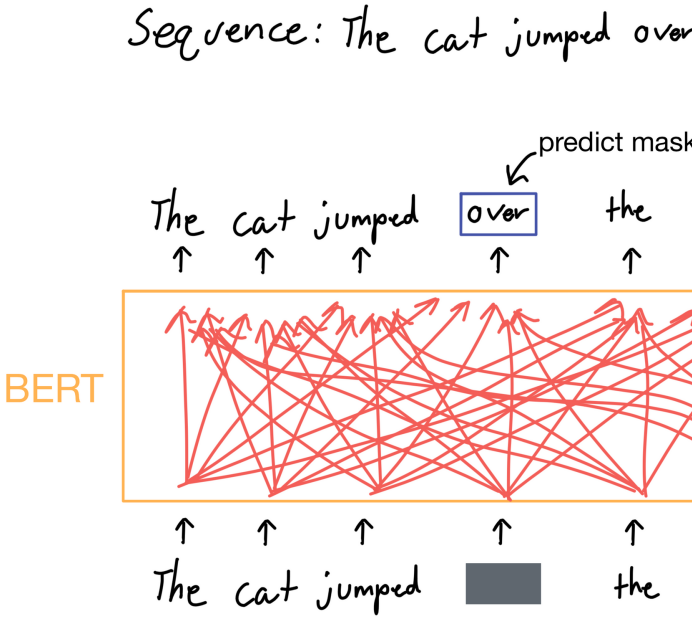
Single-mode approaches

Multi-modal approaches

Speech Representation Learning Paradigms

Generative learning

- Reconstruct input based on limited view
- Predict:
  - Future inputs
  - Masked inputs
  - Original from corrupted/noisy input



Speech Representation Learning Paradigms

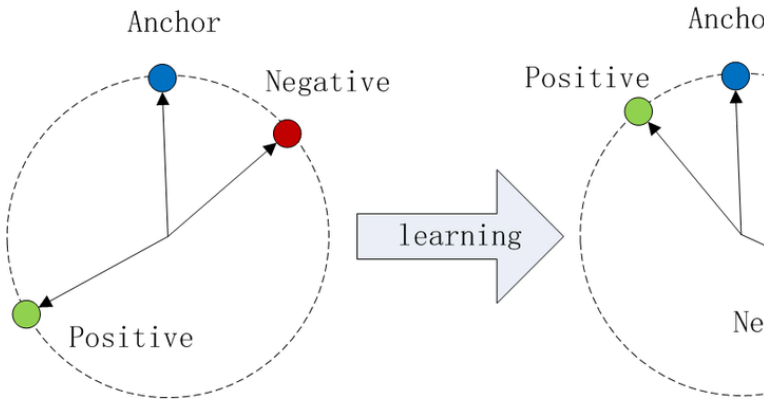
Generative learning

- Reconstruct input

Contrastive learning

- Similar samples should have similar representations

- Example loss: = anchor/target sample,
=positive,
=negatives and positive



Speech Representation Learning Paradigms

Generative learning

- Reconstruct input

Contrastive learning

- Similar samples, similar representations

Predictive learning

- Compute targets typically with another model
- Example (HuBERT)
  - Learn learning targets with e.g., k-means

Intro

Learning paradigms

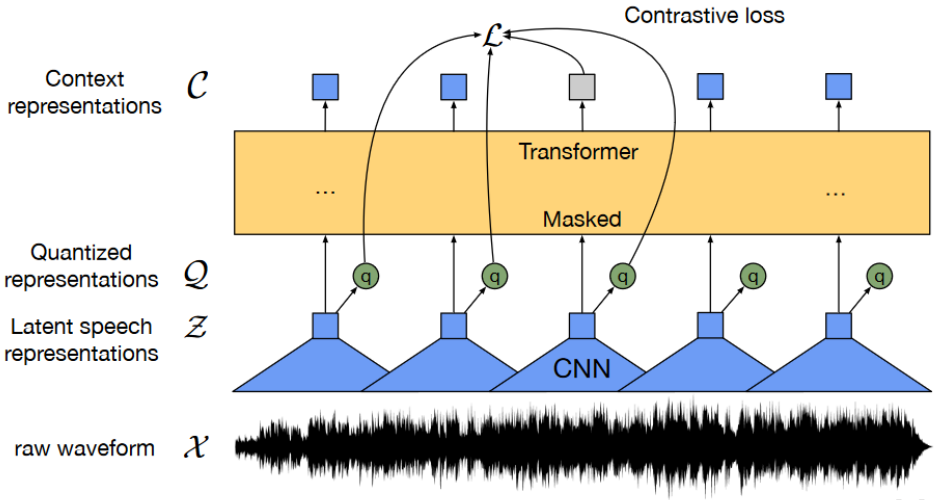
Single-mode approache

Multi-modal approaches

Single-mode approaches

wav2vec 2.0 [7]

- CNN + Transformer architecture
- Contrastive learning
  - Maximize similarity between contextualized and localized representation
  - Targets are taken only at masked timesteps



[7]

$$\mathcal{L} = -\log\left(\frac{\exp\left(s_c(h_t, q_t)\right)}{\sum_{i \in I} \exp\left(s_c(h_t, q_i)\right)}\right)$$

Single-mode approaches

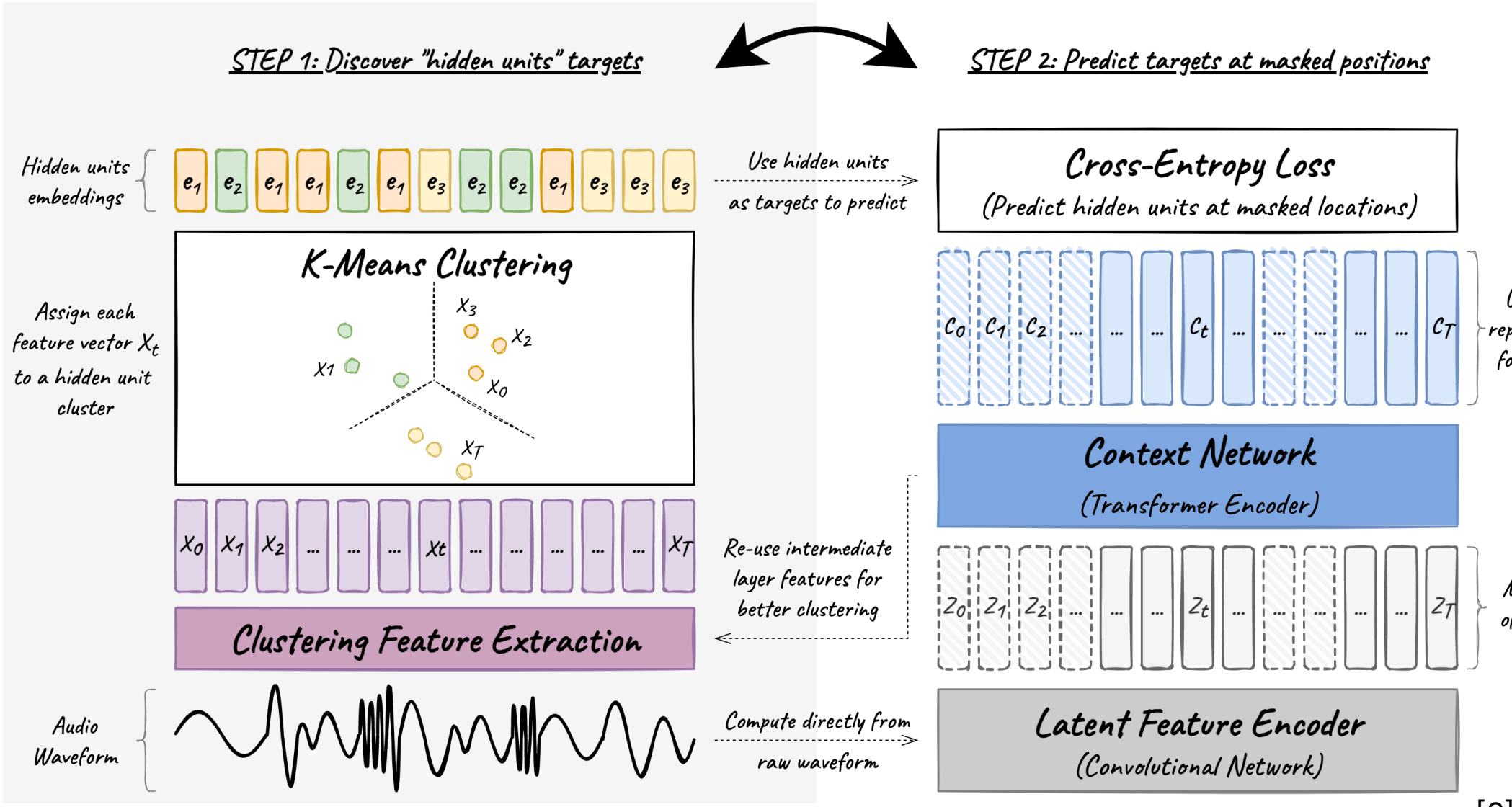
wav2vec 2.0 [7]

- Contrastive learning

HuBERT (Hidden Unit BERT)

- CNN + Transformer
- Predictive learning
  1. Cluster inputs features with k-means ✉ targets
  2. Predict targets for masked inputs

  ✉Repeat with intermediate features of HuBERT encoder


STEP 1: Discover "hidden units" targets
STEP 2: Predict targets at masked positions
Hidden units embeddings
e1 e2 e1 e1 e2 e1 e3 e2 e2 e1 e3 e3 e3
Use hidden units as targets to predict
Cross-Entropy Loss
(Predict hidden units at masked locations)
K-Means Clustering
Assign each feature vector Xt to a hidden unit cluster
X3
X2
X1
X0
XT
C0 C1 C2 ... ... Ct ... ... CT
Context Network
(Transformer Encoder)
X0 X1 X2 ... ... Xt ... ... XT
Re-use intermediate layer features for better clustering
Z0 Z1 Z2 ... ... Zt ... ... ZT
Clustering Feature Extraction
Audio Waveform
Compute directly from raw waveform
Latent Feature Encoder
(Convolutional Network)

[9]

Single-mode approaches

wav2vec 2.0 [7]

- Contrastive learning

HuBERT (Hidden Unit BERT)

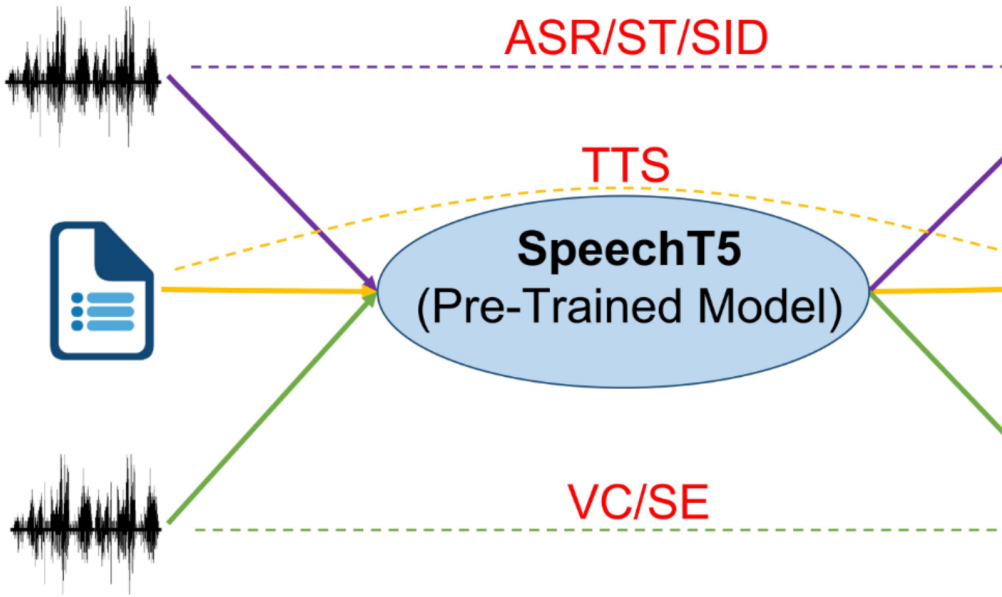
- Predictive learning

Many more… [1]

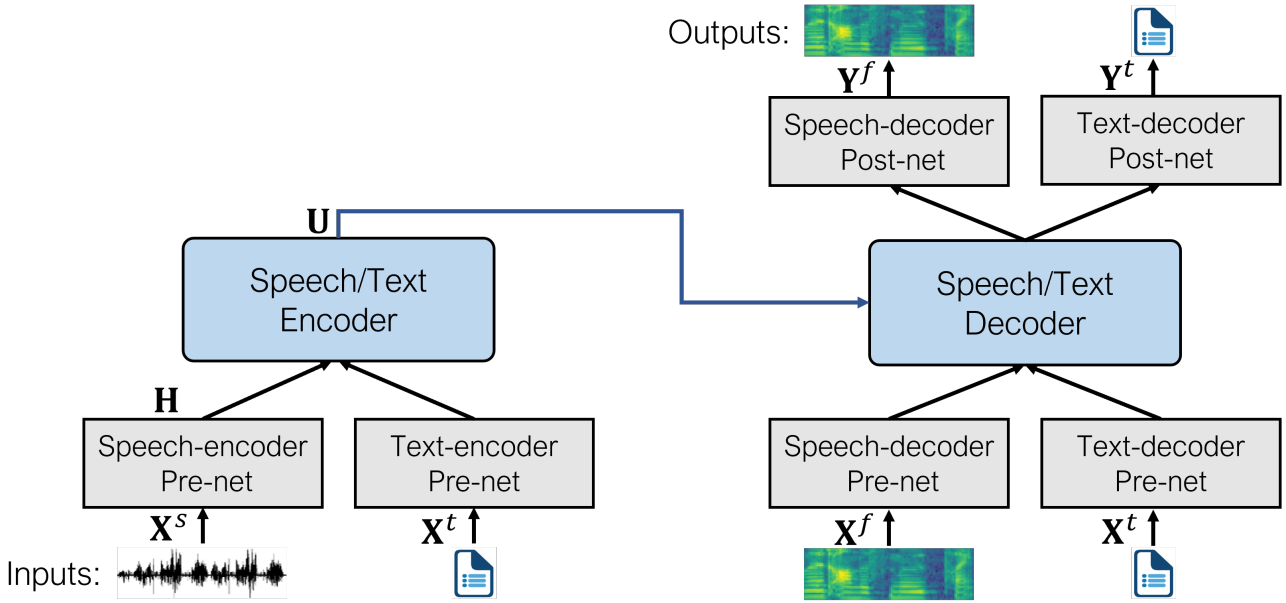
Intro
Learning paradigms
Single-mode approaches
Multi-modal approaches

SpeechT5 (10/21)

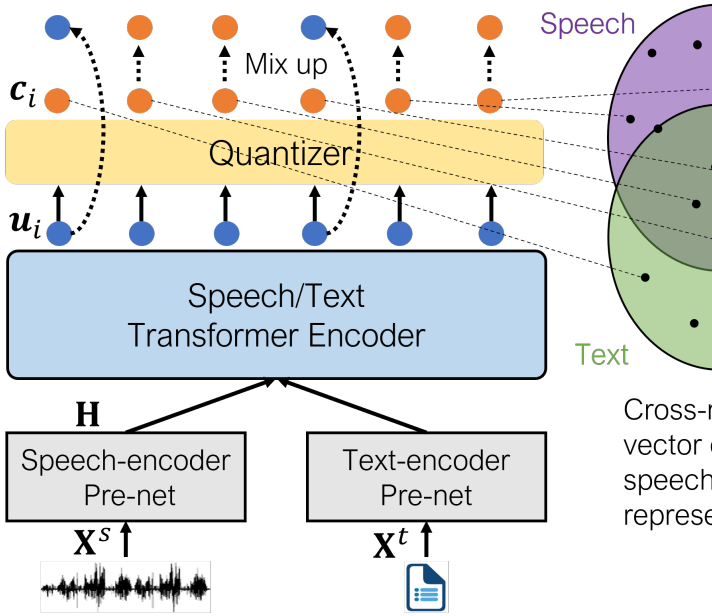
- Speech/Text representation learning
- Expansion of T5 framework
  “**T**ext-**t**o-**T**ext **T**ransfer **T**ransformer”



SpeechT5: Architecture



(a) The model architecture of SpeechT5



(b) The joint pre-training approach

SpeechT5: Training

- Learning objective (Generative):

$$\mathcal{L} = \mathcal{L}^s_{mlm} + \mathcal{L}^s_1 + \mathcal{L}^s_{bce} + \mathcal{L}^t_{mle} + \gamma \mathcal{L}_d.$$

- Pre-trained with uncorrelated, unlabeled speech and text data
- Shared representation space

SpeechT5
ASR Performance

Model	LM	dev-clean	dev-other	test-cl
wav2vec 2.0 BASE (Baevski et al., 2020)	-	3.2	8.9	3.4
Baseline (w/o CTC)	-	3.1	7.8	3.1
Baseline	-	2.8	7.6	2.8
SpeechT5 (w/o CTC)	-	2.8	7.6	3.1
SpeechT5	-	**2.5**	**7.4**	**2.7**
wav2vec 2.0 BASE (Baevski et al., 2020)	4-gram	2.0	5.9	2.6
wav2vec 2.0 BASE (Baevski et al., 2020)	Transf.	1.8	4.7	2.1
Baseline	Transf.	2.0	4.5	1.9
SpeechT5	Transf.	**1.8**	**4.3**	**1.9**

Baseline: SpeechT5 initialized by HuBERT BASE model
Dataset: Audio: LibriSpeech [5], Text: LibriSpeech LM training set (400M sentences)

SpeechT5
Ablation Study

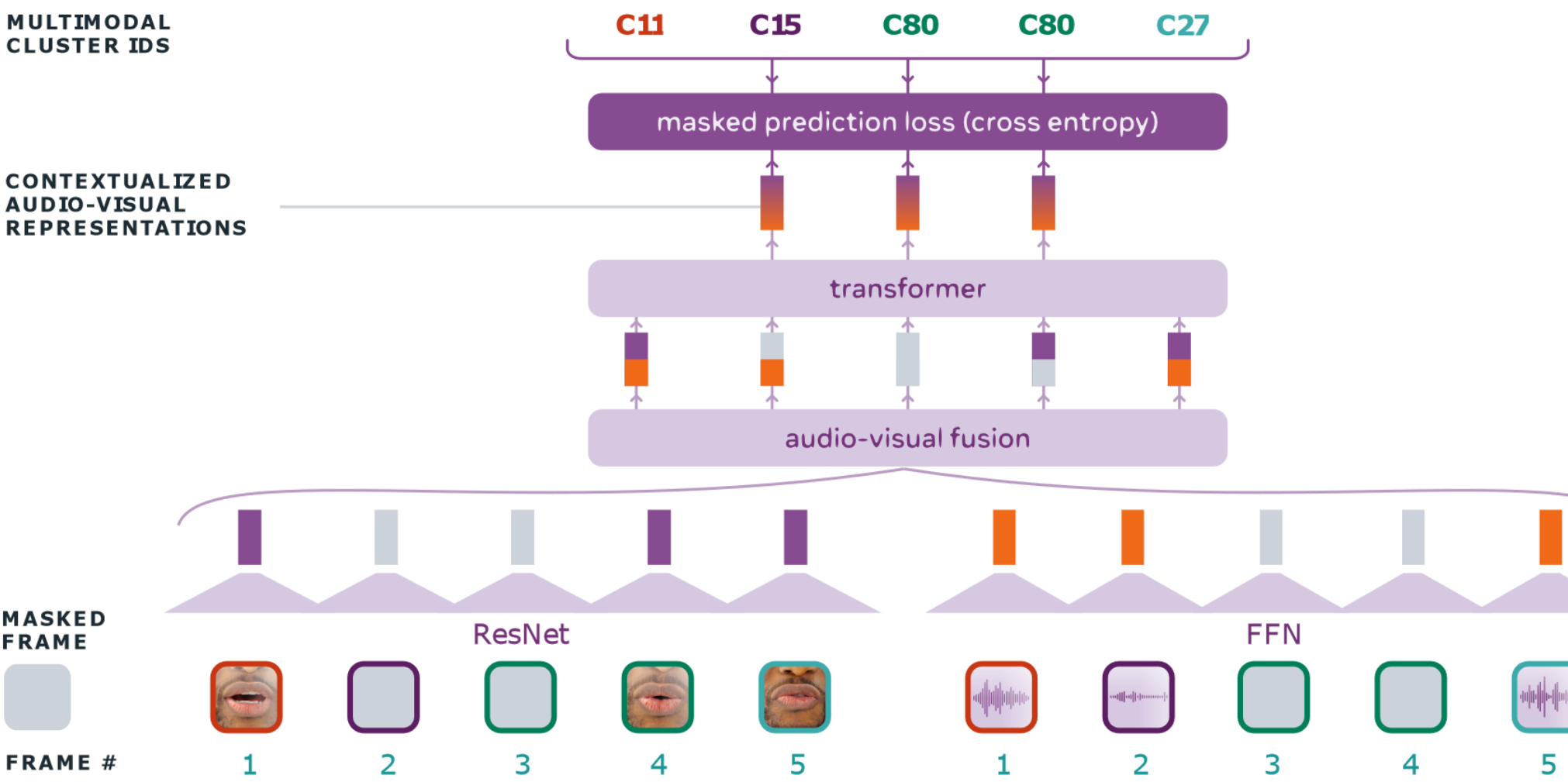
Model	ASR		VC	S
	clean	other		
SpeechT5	4.4	10.7	5.93	96.4
w/o Speech PT	-	-	6.49	38.6
w/o Text PT	5.4	12.8	6.03	95.6
w/o Joint PT	4.6	11.3	6.18	95.5
w/o $\mathcal{L}^s_{mlm}$	7.6	22.4	6.29	90.9

is cross-entropy loss for speech PT (has two losses)

AV-HuBERT (03/22)

- Audio and visual channels as input
- Based on HuBERT
  - Cluster targets from features (k-means)
  - Predict targets from masked inputs

AV-HuBERT: Architecture



AV-HuBERT: Training

- Modality dropout

$$\mathbf{f}_t^{av} = \begin{cases} \texttt{concat}(\mathbf{f}_t^a, \mathbf{f}_t^v) & \text{with } p_m \\ \texttt{concat}(\mathbf{f}_t^a, \mathbf{0}) & \text{with } (1-p_m)p_a \\ \texttt{concat}(\mathbf{0}, \mathbf{f}_t^v) & \text{with } (1-p_m)(1-p_a) \end{cases}$$

- Masking by substitution

$$\tilde{\mathbf{I}}^v_{s_i:t_i} = \mathbf{I}^{v,f}_{p_i:p_i+t_i-s_i}, \forall 1 \leq i \leq n \quad M = \{(s_i, t_i)\}_{1 \leq i \leq n}$$

- Generative and predictive learning (masked prediction / learned targets)

AV-HuBERT
Lip-Reading (VSR) Performance

Method	Backbone	Criterion	Labeled iso (hrs)	Labeled utt (hrs)	Unlabe data (h
Supervised					
Afouras et al. (2020)	CNN	CTC	157	433	-
Zhang et al. (2019b)	CNN	S2S	157	698	-
Afouras et al. (2018a)	Transformer	S2S	157	1,362	-
Xu et al. (2020)	RNN	S2S	157	433	-
Shillingford et al. (2019)	RNN	CTC	-	3,886	-
Ma et al. (2021b)	Conformer	CTC+S2S	-	433	-
Ma et al. (2021b)	Conformer	CTC+S2S	157	433	-
Makino et al. (2019)	RNN	Transducer	-	31,000	-
Semi-Supervised & Self-Supervised					
Afouras et al. (2020)	CNN	CTC	157	433	334
Ma et al. (2021a)†	Transformer-BASE	S2S	-	30	433
			-	433	1,759
Proposed (Self-Supervised & Self-Supervised + Semi-Supervised)					
AV-HuBERT	Transformer-BASE	S2S	-	30	-
			-	30	433
			-	30	1,759
			-	433	-
			-	433	433
			-	433	1,759
	Transformer-LARGE	S2S	-	30	-
			-	30	433
			-	30	1,759
			-	433	-
			-	433	433
			-	433	1,759
AV-HuBERT + Self-Training	Transformer-LARGE	S2S	-	30	1,759
			-	433	1,759

Datasets: Labeled: LRS3 [6] + Unlabeled: VoxCeleb2 [8]
Self-Training: Created pseudo-labels from voxceleb2 dataset [8] using a pre-trained ASR system

AV-HuBERT
Ablation Study (VSR)

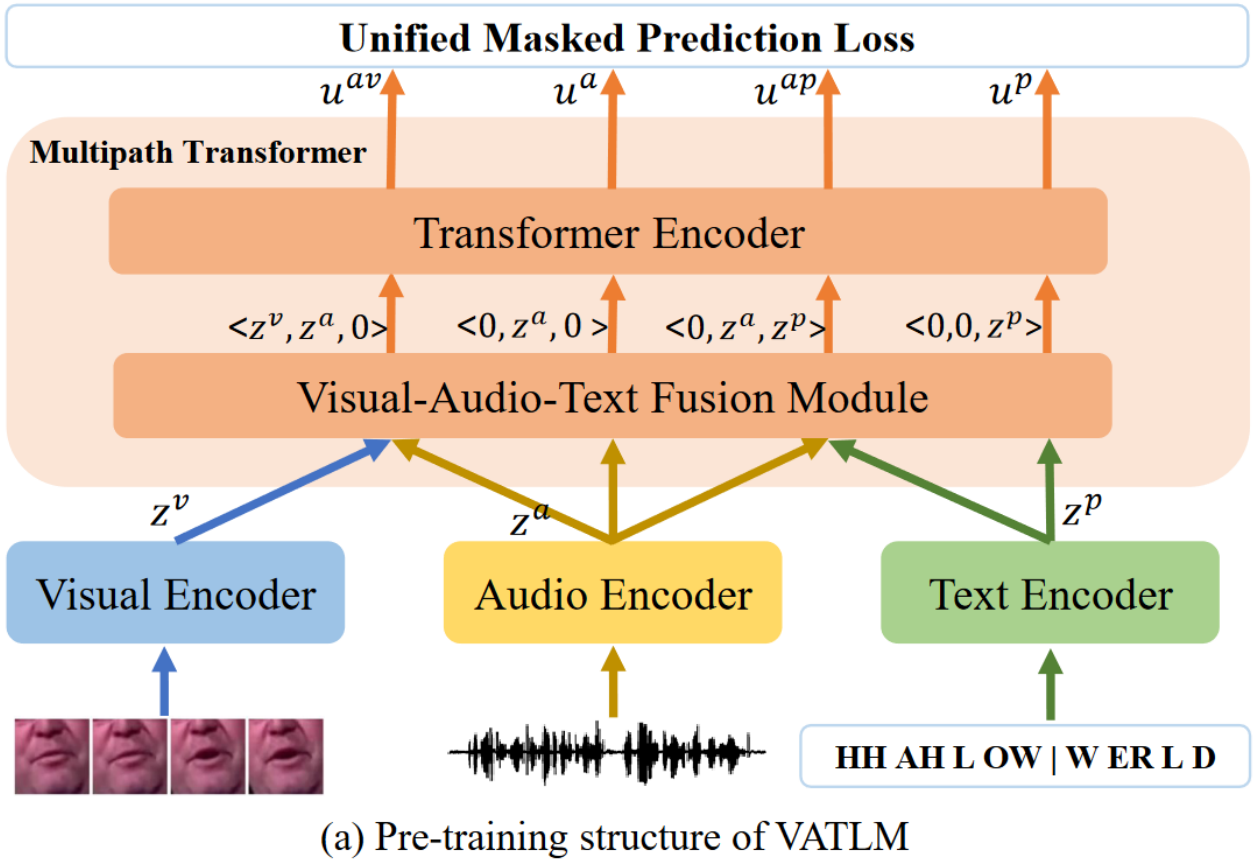
Where	Masking How	m_a	m_v	Modality Dropout p_m	p_a	Loss α	dev
Input	Sub (same, seg)	0.8	0.3	0.5	0.5	0.0	46.
	Sub (same, frm)						47.
	Sub (diff, seg)						47.
	Learned Embedding						52.
	Gauss. Noise						52.
Feature	Learned Embedding						55.
Input	Sub (same, seg)	0.8	0.3	0.5	0.5	0.0	46.
		0.8	0.8				59.
		0.3	0.3				54.
Input	Sub (same, seg)	0.8	0.3	0.5	0.5	0.0	46.
				1.0	n/a		55.
Input	Sub (same, seg)	0.8	0.3	0.5	0.5	0.0	46.
						1.0	46.

: Masking probability *a*udio and *v*isual
: Probability of *m*ulti-modal and *a*udio-only training

VAT-LM (11/22)

- **V**isual **A**udio **T**ext – **L**anguage **M**odel
- Learn representations from three channels
- Predictive/ Generative learning approach: Masked prediction task of learned targets

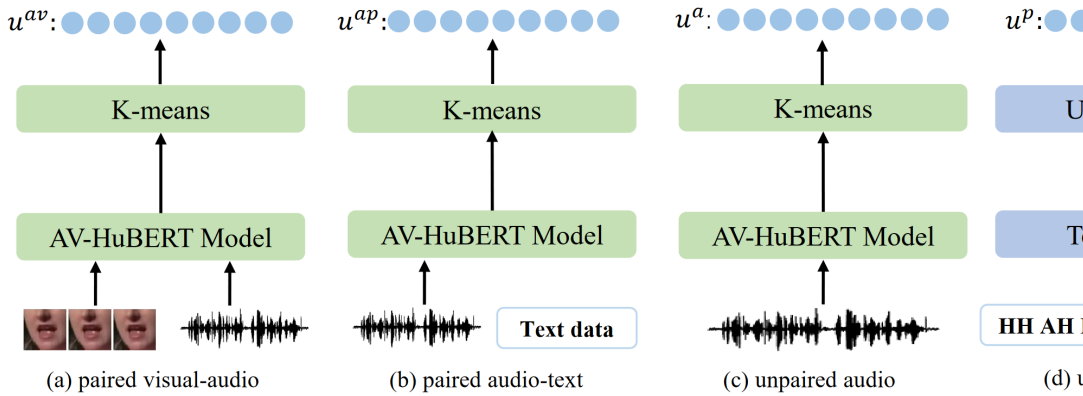
VAT-LM: Architecture



(a) Pre-training structure of VATLM

VAT-LM: Training

- Feature extraction
  - Same as AV-HuBERT
  - Text: Embedding of phonemes



- Targets from pre-trained AV-HuBERT + k-means
  - Text targets: Transcribed ASR speech data (use transcription as input)
- Task: Unified Masked Prediction (Generative)
  1. Compute targets of multimodal features
  2. Masked prediction of targets

VAT-LM
(A)VSR Performance

Method	Backbone	Criterion	Extra labeled (hrs)	Fine-tuned data (hrs)	Pre-trained AV data (hrs)	W
Supervised						
Zhang et al. [62]	CNN	CE	157	698	-	
Afouras et al. [24]	Transformer	CE	157	1362	-	
Xu et al. [18]	RNN	CE	157	433	-	
Shillingford et al. [63]	RNN	CTC	-	3886	-	
Ma et al. [25]	Conformer	CTC+CE	-	433	-	
Ma et al. [25]	Conformer	CTC+CE	157	433	-	
Prajwal et al. [44]	Transformer	CE	-	698	-	
Prajwal et al. [44]	Transformer	CE	-	2676	-	
Ma et al. [46]	Transformer	CTC+CE	-	433	-	
Ma et al. [46]	Transformer	CTC+CE	-	1459	-	
Makino et al. [19]	RNN	Transducer	-	31000	-	
Self-supervised & Semi-supervised						
Afouras et al. [41]	CNN	CTC	157	433	334	
Zhang et al. [33]	Transformer-Base	CTC	-	30	433	
Ma et al. [42]	Transformer-Base	CE	- -	30 433	433 1759	
AV-HuBERT [30], [34]	Transformer-Base	CE	- - -	30 30 433	433 1759 1759	
	Transformer-Large	CE	- -	30 433	1759 1759	
AV-HuBERT (w/ self-training) [30]	Transformer-Large	CE	- -	30 433	1759 1759	
VATLM (ours)*	Transformer-Base	CE	- - -	30 30 433	433 1759 1759	
	Transformer-Large	CE	- -	30 433	1759 1759	
VATLM (w/ self-training)*	Transformer-Large	CE	- -	30 433	1759 1759	

* In the pre-training stage, our model uses additional 3846h unpaired audio, 452h audio-text and 600M unpaired text data.

Self-training: Created pseudo-labels from voxceleb2(AV) dataset [8] using a pre-trained ASR system
Dataset: LRS3 [6]

VAT-LM
Ablation Study

#	Fine-tuned dataset	Pre-training loss	VSR WER (%)	AVS WER
1	30 LRS3	$\mathcal{L}_{total}$	48.0	3.6
2	30 LRS3	$-\mathcal{L}^{p}$	48.3	3.7
3	30 LRS3	$-\mathcal{L}^{a}$	48.3	3.9
4	30 LRS3	$-\mathcal{L}^{p} - \mathcal{L}^{a}$	49.2	4.2
5	30 LRS3	$-\mathcal{L}^{p} - \mathcal{L}^{a} - \mathcal{L}^{ap}$	51.8	4.9
6	30 LRS3 + TED-LIUM3	$-\mathcal{L}^{p} - \mathcal{L}^{a} - \mathcal{L}^{ap}$	-	4.7

TED-LIUM3: Transcribed audio-text data
Use of TED-LIUM in pre-training (#4) vs. in fine-tuning (#6)

Multi-modal Speech Representation Learning: Conclusion

Advantages

- Can outperform single-mode approaches at ASR/VSR/… [2-4]
- Often require less labeled data than single-mode approaches [4]

Challenges

- Typically, domain-specific data
- Visual Speech Recognition task is still challenging

Sources

[1] Mohamed, Abdelrahman, et al. "Self-supervised speech representation learning: A review." IEEE Journ
Selected Topics in Signal Processing (2022).

[2] Ao, Junyi, et al. "SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processi
ACL (2022).

[3] Zhu, Qiushi, et al. "Vatlm: Visual-audio-text pre-training with unified masked prediction for speech
representation learning." IEEE Transactions on Multimedia (2023).

[4] Shi, Bowen, et al. "Learning audio-visual speech representation by masked multimodal cluster predicti
ICLR (2022).

[5] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpu
based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoust
Speech and Signal Processing, pages 5206–5210. IEEE.

[6] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visu
speech recognition, 2018b. arXiv:1809.00496.

[7] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed and Michael Auli. wav2vec 2.0: A Framework for
Supervised Learning of Speech Representations, 2020. arXiv:2006.11477.

[8] J. S. Chung, A. Nagrani, and A. Zisserman, “Voxceleb2: Deep speaker recognition,” in ISCA Interspeech,
2018, pp. 1086–1090.

[9] https://jonathanbgn.com/2020/12/31/self-supervised-learning.html. Visited 01.12.2023.

wav2vec 2.0 Loss

$$\mathscr{L}_T = -\log\left(\frac{\exp\left(S_c\left(h_t, q_t\right)\right)}{\sum_{i \in I} \exp\left(S_c\left(h_t, q_i\right)\right)}\right)$$

= anchor/target sample (masked)
=unmasked (quantized) anchor sample
= Set of positive and negatives
cosine similarity

AV-HuBERT
ASR Performance

Method	Backbone	Criterion	LM	Labeled data (hrs)	Unlabeled data (hrs)
		Supervised			
Afouras et al. (2018a)	Transformer	S2S	√	1,362	-
Afouras et al. (2018a)	Transformer	CTC	√	1,362	-
Xu et al. (2020)	RNN	S2S	-	433	-
Ma et al. (2021b)	Conformer	CTC+S2S	√	433	-
		Self-Supervised			
(HuBERT) Hsu et al. (2021a) (A/MFCC→A)	Transformer-Base	S2S	-	30	433
			-	30	1,759
			-	433	1,759
	Transformer-Large	S2S	-	30	433
			-	30	1,759
			-	433	1,759
		Proposed (Self-Supervised)			
(also HuBERT) A/MFCC→AV	Transformer-Base	S2S	-	30	433
			-	30	1,759
			-	433	1,759
	Transformer-Large	S2S	-	30	433
			-	30	1,759
			-	433	1,759

Datasets: Labeled: LRS3 [6] + Unlabeled: VoxCeleb [8]
A/MFCC✉AV: HuBERT trained on AV-HuBERT targets

MFCC Features